中 文 信 息 学 报
第20卷 第4期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.20 No.4
文章编号：1003-0077（2006）04-008-08

引入标点处理的层次化汉语长句句法分析方法[1]

李幸 宗成庆

（中国科学院自动化研究所 模式识别国家重点实验室,北京 100080）

摘要：在分析汉语标点符号用法和句法功能的基础上，本文提出了一种新的面向汉语长句的层次化句法分析方法。这种方法和传统的不考虑标点符号的一遍分析方法的主要区别在于两个方面：第一，利用部分标点符号的特殊功能将复杂长句分割成子句序列，从而把整句的句法分析分成两级来进行。这种“分而治之”的策略大大降低了在传统的一遍分析方法中同时识别子句或短语之间的句法关系以及子句和短语内部成分的句法关系的困难。第二，从大规模树库中提取包含所有标点符号的语法规则和相应概率分布信息，有利于句法分析和歧义消解。实验证明我们的方法与传统的一遍图表(chart)分析方法相比，能够大大减少时间消耗和歧义边的个数，并且提高了复杂长句分析的正确率和召回率约7%。

关键词：人工智能；自然语言处理；句法分析；标点符号；层次化分析方法
中文分类号：TP391 **文献标识码**：A

A Hierarchical Parsing Approach with Punctuation Processing for Long Chinese Sentences

Xing Li and Chengqing Zong

(National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100080, China)

Abstract: Based on the analysis of the usage and the syntactic function of Chinese punctuations, this paper proposes a new hierarchical approach to parsing the long Chinese sentences. In traditional parsing approaches, the parsing procedure is performed on one-level and the punctuation marks are not specially treated. Correspondingly, in our approach the complex long Chinese sentences are broken into sub-sentences or units (say ‘units’ hereafter) by using of the punctuation marks with special functions, so that the original whole sentence is parsed unit by unit. This idea of ‘dividing-and-ruling’ greatly reduces the difficulty in the traditional parsing approaches to recognize the syntactic relationship between the sub-sentences and phrases or inside the sub-sentences or phrases. And also, in our approach the grammatical rules with punctuation marks and their probabilities are extracted from the large scale Treebank, which are very beneficial for the syntactic disambiguation. Our experimental results have shown that comparing with the traditional Chart parsing algorithm, our approach can significantly reduce the time consumption and the numbers of ambiguous edges, and about 7% of the correct rate and the recall rate have been increased in parsing the long Chinese sentences .

Key words: artificial intelligence; natural language processing; parsing; Chinese punctuations; hierarchical parsing approach

1 引言

句法分析算法的时间复杂度和句子长度密切相关，在不采用启发式策略处理的前提下，典型句法分析算法的时间复杂度都近似于句子长度的三次方[1]。（在本文中提到的句长的

收稿日期：2005-06-19 **定稿日期**：2005-06-02
基金项目：国家自然科学基金资助项目（60375018，60175012，60121302）；中科院海外学者基金资助项目（2003-1-1）
作者简介：李幸（1979-），女，硕士研究生，研究方向为自然语言处理。

概念，统指句子中包含的分词以后的词的个数，而非汉字的个数。）因此，当句子达到一定长度时，句法分析的效率问题就凸显出来。我们进一步分析发现，长句带来的问题不仅仅是时间效率上的，子句边界界定的困难和错误使得某些子句的句法关系常常被割裂，并且一个局部子句分析的失败也会导致整个长句得不到正确的句法分析树。这种情况使得现有的句法分析算法，对超过一定长度的句子句法分析的正确率和召回率呈现急剧下降的趋势。例如，文献[1]实验得到汉语句法分析的正确率当句子长度在20个词长以下时，可以得到将近80%的正确率，而当句长超过20个词时，正确率呈现急剧下降的趋势。

然而，目前对于长句的结构模式，以及句型与表意模式之间关联方式的研究，还处在刚刚起步的阶段，并且句法分析之前，句型通常是未知的，这使得我们需要考虑从其它方面入手，尝试解决这个问题。汉语中逗号、冒号和分号是三个最常用的连接简单句和短语使其成为长句的标点符号。在Chinese LDC发布的TCT 973树库[3]上的实验数据表明，在随机抽取的4431个20个词长以上的长句中，至少含有上述三种标点符号之一的长句总共有4075句，约占92%。因此，标点符号应用的普遍性使得我们研究标点符号作为连接长句中子句的唯一显式标记的用法和规律，并利用标点的作用帮助长句句法分析。

根据上述思路，本文从研究汉语标点符号在句子中的作用和规律入手，根据一些特定标点将长句切分处理，提出了一种适用于汉语长句句法分析的分层处理方法，并通过实验证明该方法在处理复杂长句的效率和正确率两方面，与传统的一遍分析方法相比较具有较大的优势。并且，从树库中提取包含标点符号的语法规则和概率信息，将其应用于句法分析过程，在一定程度上有助于句法分析的内部歧义消解。

本文的其余部分组织如下：第二部分介绍相关的研究背景和中英文标点符号研究的相关工作；第三部分比较中英文标点符号的不同，讨论了汉语长句分析的特殊困难并提出了解决问题的思路；第四部分详细介绍了我们提出的分层次处理方法；第五部分给出最终的实验结果和相应的分析；最后是本文的结束语。

2 相关研究

到目前为止，根据我们的了解，汉语标点符号在句法分析方面的研究主要集中在语言学应用方面，通常是语言学家、教育学家和编辑的研究目标。尽管标点符号是书面汉语的一个重要的组成部分，绝大多数现有的自动句法分析系统都忽略了它们的作用。在英语句法分析方面，一些与标点符号相关的研究已经开展，但对于汉语标点符号从自然语言处理角度的研究开展的很少，在我们所了解的范围内，尚没有看到相关研究的报道。造成这种现象的一个重要原因就是缺乏一个具有较好的一致性、并且不仅仅停留在直觉层面的标点符号相关理论[2]，另外一个就是汉语标点尤其是逗号的使用具有较强的随意性。

为了解决长句句法分析的困难，在英语方面，Nunberg[3]和Jones[4]等人率先开展了英语标点符号理论的研究，他们用大量理论和实验数据证明：在长句句法分析中融入标点符号信息是有效的。

由于汉语的标点符号体系是在借鉴西方语言的基础上构建的[4]，因此，在中英文标点符号之间有着很大相似性。考虑到这种相似性，我们认为，汉语标点符号应用于句法分析也应该是有帮助的。而且，我们可以借鉴英语方面的相关研究成果和经验。当然，除了相似点外，中英文标点符号在用法方面还存在着不同点。因此，专门针对汉语标点符号的特点开展汉语句法分析方法的研究是十分必要的。

Meyer[5]的工作，是第一个尝试根据语料库，从语言学角度对标点符号进行研究的。他把美式英语的标点符号分类，介绍了它们的功能。但他的分析是泛化的，并没有结合实际的应用。

Nunberg的“The Linguistics of Punctuation” [3]一书成为绝大多数后来的研究者们从句法分析的角度研究标点符号的理论基础。在他的研究工作中，提出了两级文法的概念，它们分别作用在不同的语法层级上。这两级文法分别为“lexical grammar” 和“text grammar”。其中“lexical grammar”定义了标点符号分隔开的句法成分（从句，短语）内部的句法关系，而“text

---

* 关于973树库的详细信息参见：http://www.chineseldc.org/index.htm。

grammar”定义了标点符号与其分隔开的句法成分之间的关系。

Nunberg这种把标点符号看作独立的语言学子系统，与普通的文本语法相互分离而又相互作用的方法，成为其它相关研究的基础和出发点[6]。

基于上述理论，Jones提出了其使用统一文法（integrated grammar）的方法[4]。他从经过句法分析后的语料库中提取包含标点符号的语法规则，经过规范化和归纳得到处理后的语法规则。他按标点符号的作用将其分为两类：并列标点（Conjoining punctuations）和依附标点（Adjoining punctuations）。并列标点表示并列成分之间的并列关系。依附标点的作用则认为仅仅是依附于邻近的句子成分。并且，并列标点也可以看作满足特殊依附原则的依附标点。因此，在他的理论中，所有的标点符号最终可以看作依附于邻近句法成分的，而并非句法上独立的个体。基于上述观点，他给出了一个统一的文法。

Jones的方法显示了很好的一致性，然而，他设计的文法只能覆盖所有标点现象中的一部分。实验数据显示，用该文法分析10个包含未涵盖标点语法的复杂句子时，有7个句子得不到分析结果[4]。

与Jones的方法不同，Briscoe等[7,8]把标点看作独立的句子成分，他们构建了有限从句文法（Definite Clause Grammar）的规则体系，用来描述标点和句子成分相互作用的规律。他们的实验表明：去掉句子包含的标点符号，大约8%的句子将得不到句法分析结果。

在汉语方面，Zhou qiang[9]利用标点符号来进行并列短语的自动获取。在机器翻译方面，宗成庆[10]和黄河燕等[11]利用标点符号和邻近的关系代词配合，从而把复杂句子切分成多个独立的简单句。总之，上述工作都没有从句法分析的角度对标点符号进行全面研究和分析。

3 层次化汉语长句句法分析方法的提出

3.1 中英文标点符号的异同分析

汉语中存在一些英语所没有的标点符号，这些标点符号通常具有明确的作用，因此，对句法分析具有很强的提示作用。最常用的这类标点包括顿号“、”和书名号“《》”。其中，左右书名号当中的句子成分，无论句法结构是什么，其作用必然是标示一本书的名字。顿号则取代英语中的逗号作为汉语中并列词语或者短语的分隔标记。例如，英语句子“I like to walk, skip, and run..”对应汉语的译句为“我喜欢走、跳、和跑。”显然，由于汉语中顿号的唯一作用是作为并列成分的标志，因此对于汉语句中并列成分结构的获取比英语简单。

除此之外，汉语和英语对于同样的标点使用方法也有不同，这里不再详述，请参阅[12,16]。

3.2 汉语长句句法分析的特殊困难

汉语的表意型语言特点使得汉语长句的构成方式具有和英语等西方语言不同的特点。从本质上讲，英语是一种“结构型”语言，一个完整的句法结构即表示一个完整的句子。当多个单句连接起来构成复句的时候，单句与单句之间需要显式的连接词或者短语。汉语则不同，汉语“表意型”的语言特点，使得汉语句子通常是表达一个完整意思的语言单元，这种特点在长句中表现得特别明显。因此，在汉语中存在一种独特的长句构成方式，就是一连串独立的简单句通过逗号或分号，连接成一个复杂的“句群”式的长句。这些长句内部的各个简单句是为了表意的需要而连接在一起的，它们彼此的句法结构完全是独立的，表示彼此之间逻辑关系的连接词不是必须的。因此，在很多情况下，它们之间的分隔标记仅仅是一个逗号或者分号。这类长句在汉语中称之为“流水复句”[13]。如下面一个例句：“我现已步入中年，每天挤车，搞得我精疲力尽，这种状况，直接影响我的工作，家里的孩子也没人照顾。”本文在TCT 973树库的实验数据表明，在随机抽取的4431个20个词长以上的长句当中，共有1830个流水复句，占全部长句的41.3%。而这种现象在不太规范的书面语场合出现的概率更高。

在这种情况下，连接句子中各个单句的唯一的显式标记就是逗号和分号等标点符号。而这种单句与单句之间缺乏连接词的情况，使得[10]和[11]用标点和连接词配合来切分复杂长句的方法不再适合。而对于传统的一遍分析的句法分析方法来讲，如果直接对长句进行分析

的话，识别长句内部单句的边界和分析单句内部的句法结构需要同时进行，这毫无疑问将会增加句法分析系统的处理难度，也是造成现有的分析系统处理缓慢和失败的一个重要原因。

3.3 依据标点分层的解决方案

为了解决上述困难，同时，将本文3.1节提到的书名号、顿号对于句法分析的明显提示作用有效地应用到句法分析系统当中，我们提出了依据标点符号分割长句，利用包含标点符号的文法规则进行分层次句法分析(HP)的方法。Nunberg的两级语法理论给我们的分层方法提供了理论基础[3]。

根据本文2.2节给出的Nunberg的两级文法的定义，其基本思想相当于找到一种合适大小的语言单元，这些语言单元是相对独立的，其内部的结构关系不受或者很少受周围的语言单元的影响。描述这种语言单元内部语法关系的语法“ lexical grammar”和语言单元之间关系的“text grammar”就可以结合起来，完成整个句子的句法分析。

本文采用上述“分解”的思想来进行汉语长句的分析，与基于语块的层次分析方法不同。在我们的方法中，利用部分标点为分界点把长句分割成句子单元的序列，在完成各个句子单元的结构分析之后，再合并起来分析得到完整的句子结果。而作为句子单元分界点的标点符号被定义为“分割”标点，其余标点被定义为“普通”标点。

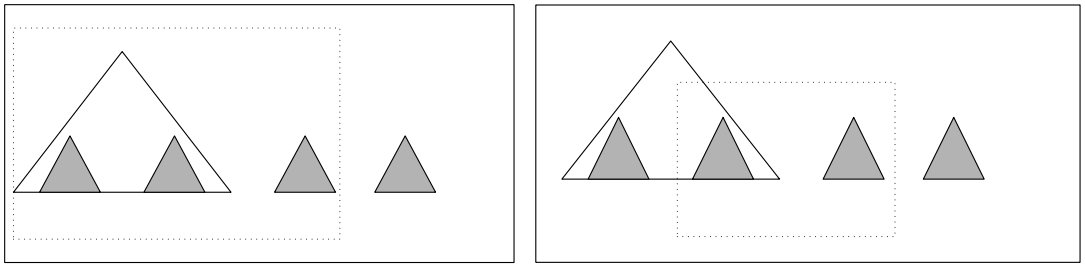
这种方法分割得到的句子单元通常是子句或者短语，因此，我们可以在第一级分析中完成对子句或短语内部的结构分析，而子句边界的获取，以及对子句或短语等彼此之间的句法关系的获取则在第二级分析完成。这种方法减少了“流水复句”以及其它类型复句句法分析的困难，这也是我们提出的HP方法的中心思想所在[15]。

4 引入标点处理的层次化长句句法分析方法

4.1 汉语标点符号的用法和分类

本文认为，可以用作“分割”作用的标点符号须满足如下条件：如果某个标点符号分隔开的子句单元，相互之间的句法关系是整体的而非局部的，换言之，也就是它们整体发生关系而非某个子句单元内部的某一成分和其它子句片断发生关系，那么，这种标点就属于“text grammar”层面，我们把这类标点定义为“分割”标点。

下图用图1来形象地表示这种定义满足的条件。其中，子图*a*所示的虚线框中的标点“P”就是“分割”标点的例子，而下图*b*所示的虚线框中的标点“P”为“普通”标点。



a: “分割”标点　　*b*: “普通”标点

图1 “分割”标点和“普通”标点

根据1995年12月13日发布的《中华人民共和国国家标准标点符号用法》[4]，通过分析可以得到，分号和冒号可以做为“分割”标点。

逗号的用法比较复杂，根据上面“国标”的定义，逗号在句子中的位置主要包括如下几种：1）句子内部主语与谓语之间；2）句子内部动词与宾语之间；3）状语和其修饰的句子之间；4）复句内各分句之间。对第四种作用，逗号充当的角色和分号的作用类似，因此可以作为“分割”标点。而对前三种情况，这些逗号分隔开的充当不同成分的短语结构可以看作是相对独立的，可以先分析各个成分内部的结构关系，在此基础上分析各个成分之间的组合结构关系。

[4] 详细信息参见教育信息网http://yywz.xhedu.sh.cn/cms/data/html/doc/2003-10/23/246 28/

那么逗号也可以当作“分割”标点。

但由于逗号的特殊性，将其定义为“分割”标点可能会造成两种错误情况：第一种是导致第一级分析时局部子句分析失败的问题，这种问题可以通过合并子句和其前后成分，进行第二级分析得到解决。第二种情况是当多个逗号分开的短语构成并列短语结构，充当一个句子成分时，此时用逗号“分割”句子，将造成这些短语被分割到其前后的句子成分当中，为了解决这种情况，我们给出一种简单的基于规则的并列成分短语的探测和合并方法，该方法在本文4.3.3节详细介绍。

4.2 语法规则的提取

本文所用的语法规则是从经过句法分析的树库中自动提取的，因此，就要求树库具备一定的规模，包含丰富的语法现象，并且标点符号用法规范。TCT 973树库就是满足上述要求的大规模树库资源。该树库的规模约为100万字，其中语料文本都选自90年代的现代汉语语料，主要分为文学、新闻、学术和应用等四类，平均句长为23.3词/句,句长在20个词长以上的句子约占一半。

首先，从树库中提取包含标点的PCFG语法规则，对包含各类标点的语法规则进行合并、概括等处理。然后，将其和由语言学分析得到的标点用法规律结合，对提取的规则进行调整。例如，仍以书名号为例，中文左右书名号之间的部分必然是一本书的名字，无论其句法类是什么，因此，可以用一条概括的语法规则描述如下：

$$NP \rightarrow 《X》 \qquad X:\{NP, VP, S, PP.....\} \tag{1}$$

在公式（1）中，X可以是所有可能的词性或者短语的类别标记。由于树库规模的限制，在树库中未出现的，而由上面分析可以得到的规则被加入到已提取的规则系统当中，并且所有此类规则的概率均为1。

除了单书名号、方括号等类似书名号的情况，其余的语法规则的概率均采用最大似然估计的方法求得。最终，所有的文法规则构成一个完整的语法规则系统。

4.3 层次句法分析方法

整个HP方法主要由三部分组成。首先对包含“分割”标点的长句进行分割，然后对分割成的各个子句单元分别独立地进行句法分析，分析得到的各个最大概率的子树根结点的词性或者短语类别标记作为第二级句法分析的输入，通过第二遍分析找到各个子句或者短语之间的结构关系，从而获得最终整句的最大概率的句法分析树。其中，在第二级句法分析之前，可以预先判断句子中是否存在并列成分短语，因此可以加入一个并列成分探测和子树合并的模块。整个算法的结构框图如图2所示：

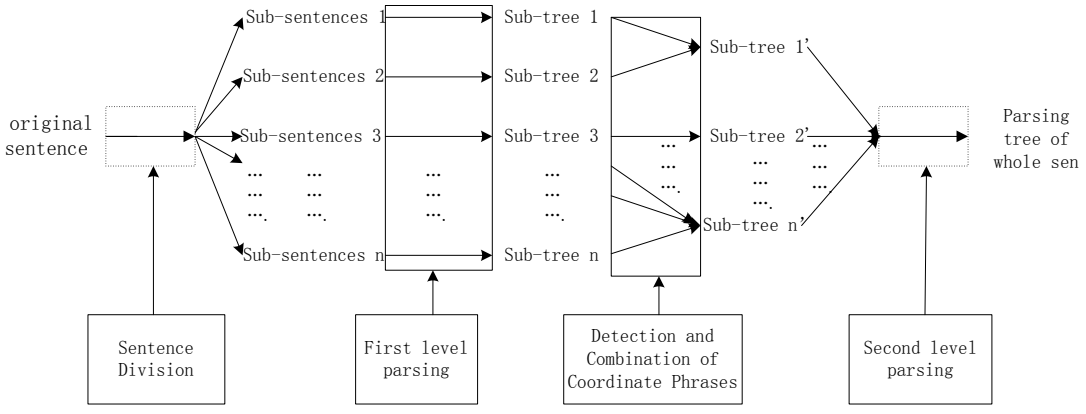


图2　HP 方法系统框架

4.3.1 长句分割

根据前面对标点符号的分类，我们用逗号、分号和冒号把长句分割为一系列子句片断。

需要注意的是，引号和破折号只具有语义上的作用，因此从句法分析上可以看作透明的[4]。

4.3.2 第一级分析

本文采用chart算法来做为分析算法，采用4.2节所述的方法建立语法规则体系。

在第一级分析当中，对经过分割得到的各个子句单元进行分析，分析的原始输入为词性串序列，用上述算法进行分析，经过韦特比(Viterbi)搜索和剪枝，对每个子句片断得到一棵最大概率的分析子树。

4.3.3 子树合并

根据4.1节的讨论，由于逗号的特殊性，将其统一定义为“分割”标点可能会造成不适当的句子分割。造成这类情形主要是由于在汉语中，当并列的短语较长时，一些句子使用逗号来替代顿号作为分隔标记，而这些并列短语充当同一句子中的一个成分。例如，句子：

我喜欢在春天去观赏桃花，在夏天去欣赏荷花，在秋天去观赏红叶，但更喜欢在冬天去欣赏雪景。 (a)

前三个动词短语在句子中作为并列的谓词短语，即第一个句子单元“我喜欢在春天去观赏桃花”中的动词短语和逗号后的动词短语是并列关系的谓语，用逗号进行分割，则会割裂这种关系，因此，可以对这种情况进行探测并且处理。

由于在第一级的句法分析中，对逗号左右的句子成分已经进行了分析，获得了逗号附近的句法结构信息，而这一步需要做的仅仅是判断逗号左右的成分是否是满足并列关系的结构完全相同的短语。

根据上面例句(a)，我们给出一个简要的并列成分短语判断过程的描述。如下图3所示，第一个逗号后的成分分析后得到动词短语(VP)，用*B*来标记。显然，*B*是由一个介词短语(PP)和一个动词短语构成的。如果第一个逗号之前存在一个最小长度的短语，并且它和*B*具有完全相同的句法结构，那么它和*B*为并列短语。显然，图中 A_2 就是这样的一个短语。其它逗号相邻成分的分析与此相似。最终得到 A_2, B 和 C为并列短语。显然，第三个逗号后的短语*D*具有和 A_2、 B 、*C*不同的结构，所以与它们不是并列关系。

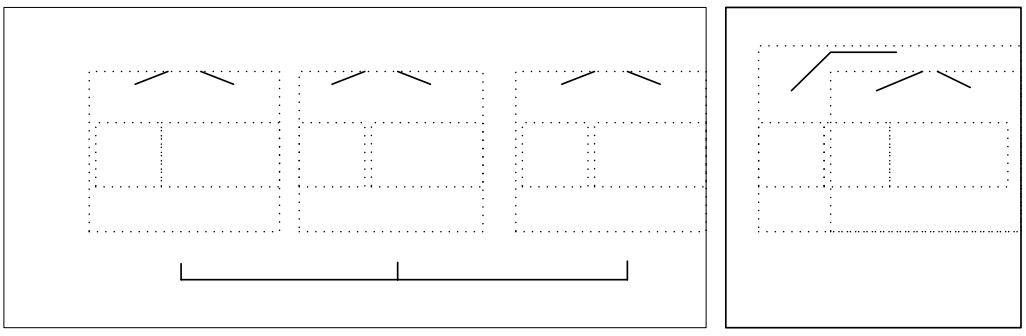


图3 例句并列结构

对这种类型的问题，我们提出了一种子树粘接操作的方法来处理，从而将并列的成分合并。如图4所示，首先把子树 A_2 和*B*、*C*合并，然后用合并后的子树 A_2' 来替换原来的 A_2 但不改变树*A*的结构。图4显示了这种子树粘接的操作过程。

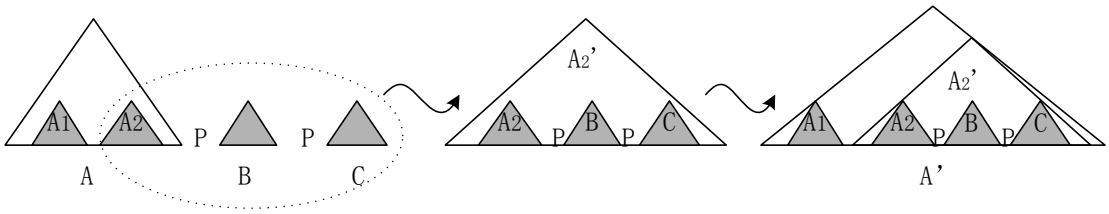


图4　子树粘接操作

以下，我们给出子树粘接操作的执行条件和规则：

$$[X,]^{+}Z[X \quad Y...Y] \Rightarrow Z[X[[X,]^{+}X] \quad Y...Y]$$

$$\mathrm{X}=\{np,vp,ap,dp\},\ Z\text{表示短语或者子句标记},\ \mathrm{Y}=*(\text{表示任何结构标记}) \tag{2}$$

$$Z[Y...Y \quad X][，X]^{+} \Rightarrow Z[Y...Y \quad X[X[，X]^{+}]]$$

$$\mathrm{X}=\{np,vp,ap,dp\},\ Z\text{表示短语或者子句标记},\ \mathrm{Y}=*(\text{表示任何结构标记}) \tag{3}$$

其中，规则(2)和(3)中的X表示的短语必须是经过我们前面判定为并列关系的短语。例如，对上面图3所示例句，判定 A_2, B 和 C为并列的VP短语，则对子句1（Sub-Sentence 1），合并前的结构为$\mathrm{S[A_1\ VP][,\ VP][,\ VP]}$，其中 A_1 的部分表示对“我喜欢”分析得到的结构标记，则采用规则(3)，合并后的结构为：$\mathrm{S[A_1\ VP[VP,\ VP,\ VP]]}$。

4.4.4 第二级分析

第二级分析所用的算法和第一级分析所用的算法是相同的，不同点在于所用的文法规则和输入串。但由于两部分文法规则有部分重叠，依靠算法自动选择即可。

输入词性串在第一级分析是各个句子单元的输入词性串序列，而第二级分析的输入则分为两种情况：第一种，第一级分析的各个子树单元都能够获得最大的概率分析树，此时第二级分析的输入则为各个分析树根结点的结构标记和分隔他们的标点符号。第二种情况是当第一级分析的某些子句分析失败时，我们仍取失败子句的原始词性序列和其前后分析成功的子树一起进行第二级的分析。

第二级分析最终输出结果是整个句子的最大概率句法分析树。

5 实验结果及分析

5.1 测试用句

由于本文的分层次分析(HP)方法的目标是利用标点符号的信息来克服复杂长句句法分析的困难，一般认为，“复杂句子”表示词项数大于等于20的句子[12]，因此，实验要针对20个词长以上的句子来进行。

首先，我们从TCT 973树库当中随机抽取8,059句，四种类型的文本各自约占四分之一左右。以上述句子为训练集，经过提取和处理得到3,795 条概率上下文无关的语法规则。然后，另外选取847个句子作为测试集，其中20个词长以下的句子被过滤掉，最终得到420个长句作为测试集。这些句子的分布情况如下表1所示：

表1 测试句分布

文体	句子数	句长（词/句）	平均句长（词/句）
文学	116	21～123	36.06
新闻	123	22～100	37.73

学术	114	21～131	39.47
应用	67	20～98	38.36
合计	420	20～131	37.84

5.2 实验结果

为了将本文的HP方法和传统的一遍扫描的方法（TP）相比较，使用如上表所示的相同的测试数据集，并且采用相同的语法规则集。

5.2.1 算法时间效率评估

在一台PC机上（奔腾4处理器，1.20GHz, 256M内存）分别运行两种方法构建的系统，时间消耗如下图5所示：

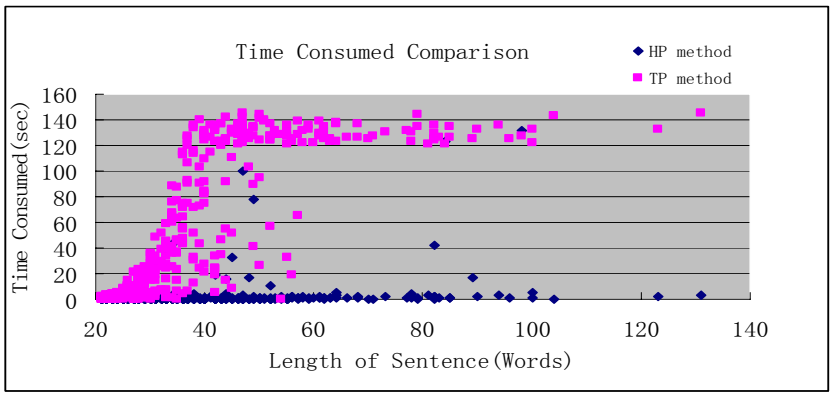
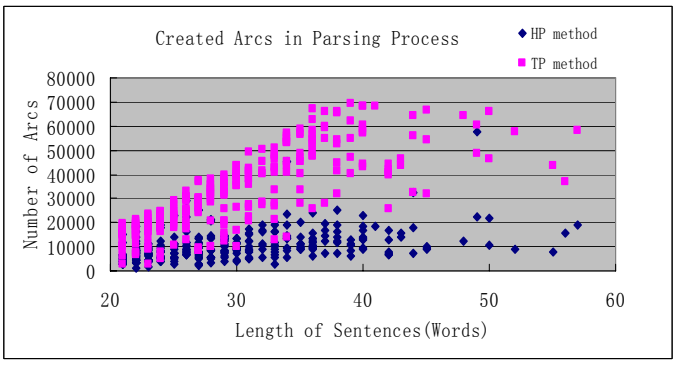


图5　时间性能

在本文的实验系统当中，设定一个句子分析时间上限值为120秒，一旦运算时间超时则退出算法循环，从而得不到最终的分析结果。传统的一遍chart分析的方法(TP)时间复杂度为句子长度的三次方。因此，如图5所示，当句子长度超过40个词时，绝大部分句子无法在给定时间内获得分析结果。而本文的HP方法由于“分解”的处理策略，句子长度增加给HP方法带来的难度增加远不明显，绝大部分句子可以在20秒时间内得到分析结果。因此，HP方法在时间性能上显然大大优于传统的TP方法。

另外，HP方法在分析过程中生成的扩展弧和歧义边数量比传统的一遍chart分析方法大大减少了，如图6所示。部分原因在于分解句子之后，在第一级分析中每个句子部分只取最优的结果进入第二级分析，相当于剪掉了局部次优的结果。剪枝的结果使得HP方法生成的中间状态的扩展弧和歧义边的数量与TP方法相比，呈现粗略的常数比的减少趋势。



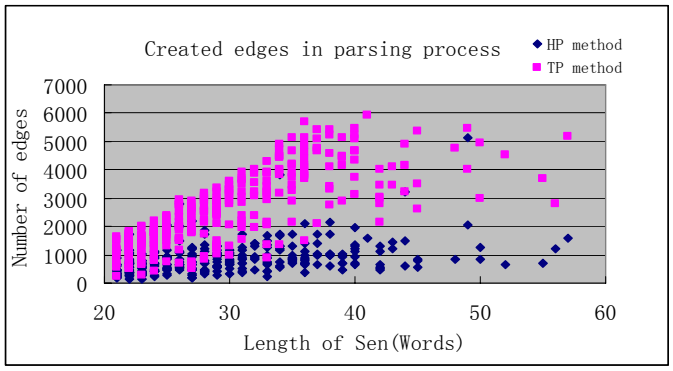


图6 扩展弧（上图）和歧义边（下图）的数量

5.2.2 算法正确率和召回率评估

首先从算法成功率方面比较两种方法，因为是开放测试，则存在无法得到完整句法分析树的情况，下面表2给出在120秒时间限制内，HP和TP方法分析失败的句子个数及百分比。

表2 失败句子分析比例

方法	句子总数	失败句子个数	百分比
TP	420	97	23.1%
HP	420	16	3.8%

显然，HP在给定时间限制内大大提高了句子分析的成功率。

除了上述分析失败的情况外，只考虑分析成功的句子，我们采用国际上广泛采用的PARSEVAL评测标准[14]算法性能进行综合的评估，测试结果如下表3所示：

表 3 采用PARSEVAL 评估的实验结果

文体	**方法**	**LP**	**LR**	**CBs**	**0CB**	**<=2CB**
文学	TP	67.31%	66.76%	6.97	19.77%	48.84%
	HP	73.57%	73.77%	3.24	40.74%	62.09%
新闻	TP	61.05%	61.69%	5.80	10.47%	34.88%
	HP	70.66%	70.58%	3.52	28.33%	61.83%
学术	TP	61.20%	60.89%	5.63	12.66%	37.97%
	HP	68.74%	68.98%	4.14	23.37%	59.10%
应用	TP	64.10%	64.61%	6.17	6.25%	27.08%
	HP	66.55%	67.81%	4.68	21.54%	50.77%
平均	TP	63.38%	63.41%	6.14	13.04%	38.46%
	HP	70.06%	70.03%	3.80	30.24%	61.01%

从上表所示的对比数据，可以看到，HP方法的正确率和召回率都高于TP将近7%，而且平均交叉括号数减少了一半，0交叉括号的百分比提高了大约一倍多，2交叉括号百分比(≤2CB)也提高了将近一倍。

另外，表中数据显示，对于不同的文体，HP方法的正确率以及相对于TP方法的改进效果存在着差别。其中，文学体裁的句子的分析正确率和召回率最高。经分析树库中文学类句子的特点发现，在116个测试句当中，“流水复句”共有97个，占了84%。上述实验结果说明HP方法对于文学体裁的句子处理效果较好。而应用类型的句子分析的正确率和召回率最低，相对于TP方法的改进效果最不明显。分析这类句子发现，相对于其他三种类型的句子，应

用类句子包含更多的嵌套名词短语和并列成分，诸如长的组织机构名和广告商品名等等，这就造成了名词短语的组合结构歧义。

需要注意的是，本文的对比方法TP算法采用了与HP算法相同的语法规则集，因此，“普通”标点符号对第一级句法分析的歧义消解作用无法在对比实验数据中反映出来。而如果去掉标点或者将句子中所有的标点统一标记为一个句法标记，实验显示，巨大的歧义性使得TP方法句法分析的失败率远远高于表2的实验结果。如此高的失败率使得我们再对TP方法分析成功的句子和HP方法进行正确率比较评测已经失去意义了。除此之外，前面提到文献[4]和[8]已经做过这方面的对比实验，因此本文不再给出这种对比实验。

6 结束语

本文从句法分析的角度出发，研究了标点符号在汉语长句分解以及句法结构歧义消解方面的作用，针对汉语长句句法分析的困难，本文提出的方法有效地利用标点符号的提示作用将长句分解，建立了一种分层次的汉语长句分析策略，并且构建了包含标点符号的语法规则体系，用于指导分析过程和句法歧义的消解。实验表明这种方法对于分析汉语长句有较好的改进效果。

从另一个角度看，这种分层的处理方法本质上是一种“分而治之”的策略，是由局部最优推导全局最优的过程，而采用标点符号实现句子分割的原则，实际上就是利用显式的标记将句子分为较大的“语块”，使得这些“语块”的内部结构分析是相对独立的。在真实自然语言条件下，由局部最优并不能保证全局最优，因此，本文的方法是在可能牺牲长句中部分单句结构分析正确率的前提下，换取其余短语和单句分析的正确率和召回率，从而提高整句分析的效率和性能。结合汉语“流水复句”普遍的现象，这种方法的处理效率和正确率还是有较大优势的。

在下一步的工作当中，要加强标点符号对于长句构成的规律研究，长句的结构模式研究以及句型与表意模式之间关联方式的研究。除此之外，借鉴文献[9]的方法预先进行短语识别，尤其是包含标点的并列短语识别，借鉴文献[10]和[11]的方法，对包含连接词的情况用标点符号配合连词进行从句识别等，以提高分析系统的性能。

参考文献:

[1] 张艳，汉语句法分析的理论、方法的研究及其应用[D]， 中国科学院自动化所，2003

[2] Jones Bernard, Towards a Syntactic Account of Punctuation[A]. In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)[C], Copenhagen, Denmark, August,1996

[3] Geoffrey Nunberg. The Linguistics of Punctuation[M]. CSLI Lecture Notes, No. 18, Stanford CA, 1990

[4] Jones Bernard. What’s the Point? A (Computational) Theory of Punctuations[D]. PhD thesis, Centre for Cognitive Science, Universito of Edinburgh, Edinburgh, UK(1997)

[5] Charles Meyer. A Linguistic Study of American Punctuation[M]. Peter Lang: New York. 1987.

[6] B. Say and V. Akman, Current approaches to punctuation in computational linguistics[J], Computers and the Humanities, vol. 30 (1997) 457-469

[7] Edward Briscoe. The Syntax and Semantics of Punctuation and its Use in Interpretation[A]. In Proceedings of the ACL/SIGPARSE International Meeting on Punctuation in Computational Linguistics[C], Santa Cruz, California. (1996) 1－7.

[8] Edward Briscoe and John Carroll. Developing and Evaluating a Probabilistic LR Parser of Part-of-Speech and Punctuation Labels[A]. In Proceedings of the ACL/SIGPARSE 4th International Workshop on Parsing Technologies[C], Prague, Czech Republic. (1995) 48－58

[9] Zhou Qiang. The Chunk Parsing Algorithm for Chinese Language[A]. In Proceedings of JSCL’99[C],

(1999) 242-247
[10] 宗成庆，张玉洁，山本和英，坂本仁，白井谕，口语自动翻译系统中的汉语语句改写[A], 中文计算国际会议（ICCC）论文集[C].2001,新加坡. 第 395-401页.
[11] 黄河燕，陈肇雄. 基于多策略分析的复杂长句翻译处理算法[J], 中文信息学报.2002,16(3).-1-7
[12] 李幸. 汉语句法分析方法研究， 中国科学院自动化所[D]，2005
[13] 周强. 汉语句法树库标注体系[J],中文信息学报.2004,18(4).1-8
[14] E.Charniak, Statistical parsing with a context-free grammar and word statistics[A]. In Proc of AAAI’97[C], 1997.
[15] Xing Li, Chengqing Zong. A Hierarchical Parsing Approach with Punctuation Processing for Long Complex Chinese Sentences[A]. In *Companion Volume to the Proceedings of Conference including Posters/Demos and Tutorial Abstracts, IJCNLP2005,* Jeju Island, Korea, October 11-13, 2005. Pages 9-14.